Linguistique de corpus et terminologie

Anne Condamines

Anne Condamines
CNRS, Equipe de Recherche en Syntaxe et Sémantique, Toulouse

LINGUISTIQUE DE CORPUS ET TERMINOLOGIE

1. Introduction

La terminologie n'a pas toujours fait très bon ménage avec les textes. La tradition wüstérienne[1] a même mis en garde contre l'utilisation de productions réelles pour constituer des terminologies. Ce n'est que récemment, sous la pression de différents paramètres, que la constitution de terminologie à partir de textes a pris un essor considérable. Une des conséquences de cette évolution est que la terminologie en tant que discipline scientifique s'est rapprochée de la linguistique. En effet, la linguistique elle-même se trouve à un tournant majeur de son histoire : les corpus sont maintenant facilement disponibles (même s'il convient de s'interroger sur cette facilité d'accès, notamment sur l'internet) et des outils pour les interroger sont également accessibles. La prise en compte des corpus vient ainsi interroger la linguistique dans de nombreux domaines: morphologie, syntaxe, discours, sémantique…
La terminologie textuelle a ainsi émergé au moment même où la linguistique de corpus se développait, en particulier la sémantique de corpus. L'objectif de cet article est de montrer que cette évolution conjointe conduit certainement à des interrogations croisées qui alimentent la réflexion de chacun des points de vue. La terminologie textuelle pourrait ainsi contribuer à éclairer beaucoup de phénomènes qui sous-tendent la sémantique de corpus, phénomènes qui ne peuvent réellement être compris que si l'on prend en compte l'objectif, théorique et/ou appliqué de l'interprétation.

2. Linguistiques de corpus

Le pluriel du titre de ce paragraphe[2] signale combien ce terme « linguistique de corpus » peut recouvrir de diversité. Pendant plusieurs années, il a surtout été utilisé par la communauté TAL (Traitement Automatique de la Langue) et il est probable que cette communauté, en proposant des outils et en ouvrant des portes sur des besoins « sociétaux » a offert de nouvelles perspectives à une linguistique qui s'appuyait sur des méthodes et des objectifs plus classiques. Par ailleurs, une partie de la linguistique a une tradition d'analyse de corpus qui reste très vivace. Il convient de faire un état des lieux sur les travaux existants et sur leurs complémentarités.

---

[1] Wüster est un ingénieur qui a publié dans les années 1930 un « dictionnaire de la machine outil » et qui est considéré comme le père de la théorie terminologique (en tout cas dans sa version normalisatrice).
[2] D'ailleurs emprunté au titre d'un ouvrage de B.Haber et al. (Habert et al., 1997)

2.1 Linguistique de corpus et informatique

2.1.1.Linguistique de corpus et TAL

Si l'on compare avec les travaux menés il y a une vingtaine d'années, qui consistaient à valider la cohérence d'une théorie élaborée sur des bases introspectives, les corpus sont désormais devenus le principal matériau du TAL (Traitement automatique de la langue). Dans cette perspective, il s'agit majoritairement de traiter de grandes quantités de données textuelles sur support électronique ; la linguistique de corpus est ainsi souvent considérée comme relevant majoritairement d'une perspective TAL :

> « Over the last three decades the compilation and analysis of corpora stored in computerized databases has led to a new scholarly enterprise known as corpus linguistics » (Kennedy, 1998).

Les objectifs de l'analyse de corpus peuvent être très différents : acquisition de connaissances morphologiques, syntaxiques ou sémantiques pour améliorer les performances des outils (l'acquisition à partir de corpus vient alors suppléer ou compléter l'approche introspective), extraction d'informations (recherche d'informations dont la nature est prédéfinie ; il s'agit alors de « remplir » des formulaires automatiquement) ; recherche d'information (retrouver les documents pertinents sur un sujet donné), système de question-réponse (il s'agit non seulement de trouver le document pertinent mais aussi de trouver la bonne réponse à une question donnée), traduction assistée par ordinateur (ce qui pose le problème spécifique de l'alignement de corpus), veille scientifique, etc.
L'un des axes qui suscitent le plus de travaux, particulièrement en France, est celui de la construction de terminologies à partir de corpus ; ce thème donne lieu à des numéros de revues (par exemple, le numéro 43-1 de la revue TAL, Hamon et Nazarenko, 2002) ou des colloques. Ce besoin en données terminologiques est apparu très nettement dans les entreprises qui doivent gérer une documentation considérable, en lien avec la création, le développement et la maintenance d'objets manufacturés[3]. La principale ressource qui alimente les outils de GED (gestion électronique de documents) est constituée par la terminologie propre au domaine couvert, voire, à l'entreprise concernée. Cette thématique permet d'établir des ponts avec un autre champ de l'informatique, celui de l'ingénierie des connaissances.

2.1.2 Linguistique de corpus et ingénierie des connaissances

L'objet majeur de l'ingénierie des connaissances concerne la constitution d'outils pouvant assister l'homme dans son raisonnement. Il s'agit d'élaborer des systèmes qui représentent la connaissance au plus près de la façon dont elle se manifeste, c'est-à-dire en utilisant les éléments langagiers propres au domaine couvert par l'outil. La plupart

[3] On considère par exemple que la documentation en volume papier d'un avion contiendrait à l'intérieur de l'avion !

des systèmes mettent en œuvre une représentation de type relationnel qui s'inscrit dans une parenté revendiquée avec les réseaux sémantiques de Quillan[4], le système le plus couramment utilisé étant certainement les graphes conceptuels de Sowa. Ces représentations relationnelles se présentent ainsi sous la forme de nœuds reliés par des arcs, les premiers étant généralement étiquetés par des noms et les seconds par des formes prédicatives (noms ou verbes). Ces modes de représentations relationnelles sont appelées ontologies en ingénierie des connaissances. Deux courants majeurs existent sur la façon de constituer les ontologies. L'un envisage la possibilité de les créer par domaine, avec un fort pouvoir de réutilisation (en tout cas supposé tel), le plus souvent sur la base du recours à la connaissance des experts de ce domaine. Un autre courant, particulièrement bien représenté en France, tient pour nécessaire le recours à des usages réels dans l'entreprise concernée et considère que les ressources terminologiques doivent être construites pour un objectif déterminé (Bachimont, 2000). Ce débat rejoint une problématique de la terminologie. Il est évident que les ontologies sont très proches des réseaux terminologiques et cette parenté n'a pas échappé à beaucoup de chercheurs. Ainsi, dès le début des années 1990, des projets interdisciplinaires autour de la constitution de « bases de connaissances terminologiques » ont vu le jour à peu près simultanément dans différentes parties du monde, par exemple au Canada (Meyer et al., 1992), à l'Université de Surrey (Ahmad et al., 1992) ou encore à Toulouse (Condamines et al., 1993). Souvent considérée comme « symbiotique » (pour reprendre un terme proposé par Skuce et Meyer), la relation entre terminologie et ingénierie des connaissances a évolué pour permettre à présent à chaque discipline de se situer et d'examiner les apports de l'une à l'autre. Pour la terminologie, la réflexion s'est nettement orientée vers le mode de prise en compte des textes et vers la définition d'une terminologie textuelle.

2.1.3 Outils d'analyse de corpus pour la constitution de terminologies

Soutenue par une demande sociétale forte, la définition d'outils d'aide à la constitution de terminologie à partir de textes est un des domaines les plus productifs du TAL. Ces outils visent à proposer des termes-candidats ou des relations-candidates. À charge pour l'utilisateur de sélectionner ceux de ces candidats qui lui semblent pertinents. Ces outils reposent sur deux grands principes qui révèlent deux façons de concevoir le fonctionnement de la langue. Dans le premier type d'outils, tout texte est considéré comme la mise en œuvre d'un système très stable. Ainsi, les termes sont considérés comme respectant des patrons récurrents (par exemple, *Nadj de N*). Quant aux relations, on considère que des marqueurs, identifiables par introspection, les mettent en œuvre de manière régulière dans les textes, par exemple *tous les N1 sauf N2* permet de repérer une hyperonymie entre N2 et N1 comme dans *Paul aime toutes les fleurs sauf les roses*. Le second type d'outil considère au contraire qu'on ne peut pas prévoir tous les phénomènes langagiers qui

[4] En 1968, Quillian a proposé de représenter sous forme de « réseaux sémantiques » (mots reliés entre eux) les phénomènes d'association dans la mémoire humaines

peuvent apparaître dans les textes et qu'il faut s'attendre à découvrir des éléments (formes de termes ou marqueurs de relation en l'occurrence) qui n'auraient pas pu être prédits par introspection. Ces outils s'inscrivent plutôt dans une tradition distributionnelle : c'est la récurrence des contextes et des distributions qui fait sens[5].
Un autre type d'outil, plus généraliste, permet une exploration du texte moins assistée mais très utile pour repérer des régularités. Il s'agit de concordanciers, plus ou moins élaborés (ils contiennent ou pas un catégoriseur grammatical), assez facilement accessibles sur internet.
On le voit, les perspectives en termes de besoins, d'outils, de méthodes en lien avec l'analyse de corpus, en particulier lorsqu'elle est appliquée à la terminologie, sont très importants.
Qu'en est-il d'une vision plus linguistique sur ce même sujet ?

2.2. Linguistique et corpus

La prise en compte de corpus en linguistique n'est évidemment pas un phénomène nouveau. (Rastier, 2001) par exemple montre bien que l'analyse de corpus est restée très vivace même pendant la période où le structuralisme puis le générationalisme ont éloigné l'étude de corpus réels de la problématique linguistique. Le mode de prise en compte des corpus peut constituer une grille de lecture des différentes approches en linguistique. Trois types de points de vue sur les corpus peuvent ainsi être dégagés : les corpus ne sont pas pris en compte, le corpus considéré est un corpus « introspectif », le corpus constitue la référence à l'analyse menée.

2.2.1 Aucun corpus n'est pris en compte

Mener une étude sur un corpus oblige à une confrontation avec la réalité des usages, ce qui ne va pas sans poser de questions à une linguistique qui voudrait s'inscrire dans une perspective scientifique et, pour cela, décrire un système stable, c'est-à-dire un système dont on maîtrise les éventuelles variations. En effet, les textes ne sont pas seulement des attestations de la mise en œuvre d'un système ; ils s'inscrivent nécessairement dans une situation particulière, qui engage des locuteurs réels et qui se caractérise par une certaine fluctuation par rapport à la norme. Une des façons de contourner cette difficulté consiste à ne pas avoir recours à des productions réelles mais, au contraire, à se donner un objet déconnecté de tout contexte afin d'établir la distance qui permettra à l'analyste de repérer les régularités inhérentes au système. C'est le choix qui sous-tend le structuralisme et le générationalisme, le premier pour des raisons essentiellement méthodologiques, le second parce que la langue est considérée comme relevant de facultés psychologiques innées et universelles.
On ne peut que constater que ces deux courants ont permis une évolution majeure dans la connaissance du fonctionnement linguistique ; le structuralisme a en particulier introduit une rupture avec une vision référentielle qui, jusqu'au début du XXe siècle biaisait considérablement

[5] Pour une description des outils de TAL pour la terminologie textuelle, voir (Bourigault et Jacquemin, 2000)

les études. Toutefois, les limites de ces approches semblent maintenant atteintes, à la fois parce que la variation fait irruption dans la problématique théorique de la linguistique et parce que la demande sociétale de résultats d'analyse de corpus est très importante.

2.2.2 Recours à un corpus «introspectif »

Beaucoup de linguistes, conscients de la nécessité de moduler les descriptions, essaient de modéliser les phénomènes de variation sur la base du recours à leur propre intuition. Leurs propres attestations constituent ainsi une sorte de corpus dont ils essaient de décrire les régularités. La plupart des chercheurs en syntaxe travaillent de cette manière mais aussi la plupart des chercheurs en sémantique lexicale (Cruse, 1986), (Kleiber, 1999) et beaucoup de chercheurs en énonciation (Ducrot, 1980).
Les résultats obtenus présentent bien sûr des intérêts : les analyses sont souvent très fines et appuyées sur de très nombreux exemples qui, pour être forgés, n'en sont pas moins acceptables. Ce qui fait problème dans ce type d'approche relève de deux ordres.
1) Les exemples sont proposés dans le cadre du test linguistique ; ainsi, ils ne sont pas irrecevables mais ils sont coupés de toute situation langagière réelle, c'est-à-dire d'une situation qui est d'abord une situation d'échange. Par ailleurs, comme le soulèvent de nombreux linguistes (Corbin, 1980) (Auroux, 1998), cette méthode accorde une grande importance au jugement du linguiste qui a tendance à ériger en règle générale sa propre acceptation des phénomènes. Tout linguiste a pourtant fait l'expérience d'entendre ou de lire un fait langagier que, quelques jours auparavant, il avait jugé comme impossible.
2) Ces exemples sont tous mis sur le même plan ; ainsi, des phénomènes rares sont considérés au même titre que des phénomènes fréquents. Cette situation s'avère particulièrement problématique lorsque l'on prend en compte des textes réels dans lesquels la répartition chiffrée des phénomènes à l'intérieur de chaque texte ou bien d'un texte à l'autre, prend un relief tout à fait significatif et participe à la construction du sens.
De plus en plus de linguistes essaient de palier ces difficultés en recourant à des attestations réelles, soit à travers l'utilisation de corpus constitués ou bien en faisant des recherches sur l'internet. Mais dans un cas comme dans l'autre, on considère souvent ces attestations comme autant de manifestations du système langagier sans tenir compte de leur origine. Dans le cas de l'utilisation de Frantext par exemple, le risque est grand de proposer des généralisations à partir d'une étude qui ne se base que sur des textes littéraires, souvent du XVIIIe ou XIXe (les plus représentés dans Frantext) qui constituent pourtant un usage langagier bien particulier. Quant à l'internet, pour qui veut essayer de comprendre la variation linguistique, il constitue une véritable énigme tant les variations de genres semblent importantes et pour l'instant, incontrôlables.

2.2.3. Corpus comme référence

L'utilisation des corpus en linguistique est loin d'être un phénomène nouveau. Cependant, si les corpus permettent de mieux rendre compte du phénomène de la variation, ils sont loin de résoudre toutes les questions qu'elle pose. Toute la problématique de la linguistique est traversée par un questionnement qui s'établit entre la nécessité de définir ce qui fait système, qui est stable et ce qui peut varier, collectivement ou individuellement, parfois jusqu'à l'infini. Dès que l'on quitte le domaine idéal du parfaitement stable, on est confronté au phénomène de la variation et à la nécessité d'en définir les contours. La prise en compte des corpus permet de poser les bases d'une réflexion sur cette problématique mais, dans l'état actuel des recherches, elle est loin d'avoir donné des réponses définitives. Dans une première approximation, on peut considérer deux points de vue selon le mode de prise en compte des corpus : l'un qui considère le corpus comme manifestant la compétence langagière et qui permet d'étudier la langue, l'autre qui considère le corpus comme la référence d'une étude particulière, les résultats ne concernant que ce corpus particulier.

Corpus comme représentatif de la compétence des locuteurs

Trois grands domaines sont concernés par la description d'une langue à partir de corpus puisque c'est de cela qu'il s'agit : la lexicologie (par exemple, Sinclair, 1995), la description de la grammaire, et enfin l'apprentissage d'une langue étrangère. L'utilisation des corpus pour ce type de perspective est nettement plus développée dans la tradition anglo-saxonne que dans la tradition francophone. On la retrouve cependant pour le français pour la constitution du *Trésor de la Langue* Française (même si sa construction est bien moins systématisée que celle du Cobuild) ou encore dans la mise en place du « français fondamental », destiné à l'apprentissage du français par des étrangers et élaboré à partir d'un corpus d'extraits attestés à l'oral ou à l'écrit.

Ce type de projets, qui a le mérite de prendre en compte la réalité des usages (de certains usages) posent tous la même question de la représentativité. En effet, s'il s'agit de décrire le fonctionnement tel qu'il se manifeste dans les corpus étudiés, il faut donc que ces corpus, puisqu'ils ne peuvent rendre compte de tous les usages, soient au moins représentatifs de tous ces usages. Or, non seulement, nul n'a les moyens de vérifier cette représentativité mais, et c'est plus ennuyeux, on voit nécessairement réapparaître l'introspection, que l'on pensait évacuer par le recours aux corpus. Des projets d'envergure essaient de baser le rapprochement des textes sur des similitudes linguistiques (Biber, 1988) ; cela suppose de distinguer une caractérisation réalisée *a priori* et une caractérisation réalisée sur la base de régularités linguistiques avérées.

Ces méthodes semblent prometteuses mais elles ne suppriment pas, pour l'instant, l'étape de définition intuitive de genres. C'est en effet sur des bases introspectives que sont définis les différents registres : discours journalistiques, textes littéraires, lettres…

Le recours aux corpus permet ainsi de se rapprocher de l'usage réel de la langue mais il ne parvient pas à constituer un objet définitivement maîtrisé.

Corpus comme objet d'étude

Certaines disciplines considèrent qu'une fois élaboré, le corpus constitue la référence à leurs travaux ; ce sont en quelque sorte les régularités propres à ce corpus qu'il faut mettre au jour. Il faut dans tous les cas que le corpus soit construit d'une manière cohérente, soit qu'il émane d'un groupe de locuteurs identifié *a priori* comme c'est le cas dans la sociolinguistique, l'analyse de discours ou même la théorie des sous-langages (Dachelet, 1994), soit qu'il soit constitué dans une perspective particulière comme dans le TAL ou la terminologie textuelle (cf. 3). En limitant la portée des résultats aux corpus auxquelles elles s'intéressent, ces disciplines ont le mérite de percevoir les limites de leur approche. Mais, en s'interrogeant peu sur les modes d'élargissement de ces résultats, elles laissent de côté des questions fondamentales pour la linguistique, qui permettent d'expliquer le fonctionnement même de la langue par la mise en œuvre de connaissances partagées (en tout cas nécessairement supposées partagées). Lors de ces analyses de corpus, quelles connaissances, déjà présentes, sont convoquées et quelles connaissances construites pourraient être réutilisées lors de nouvelles analyses ?

Ancrées dans la réalité des usages, ces approches à partir de corpus font émerger un élément majeur, particulièrement pour la sémantique. En effet, la plupart d'entre elles considèrent que le sens n'est pas un donné mais un construit et que le corpus est soumis à une interprétation. Dans une telle perspective, la question qui se pose est de savoir quels sont les modes possibles de contrôle de l'interprétation.

Dans ce cadre général qui voit de nombreux courants de la linguistique s'intéresser aux corpus, comment peut-on situer la terminologie textuelle et en quoi peut-elle éclairer des problématiques comme la systématisation des résultats obtenus sur un corpus ? C'est ce dont il va être question dans la partie 3.

3. Terminologie et textes

3.1. Avènement d'une terminologie textuelle

Dans la vision wustérienne, la prise en compte des usages manifestés dans les textes ne peut être la base de la constitution de terminologies. En effet, le discours, dans ses possibilités créatrices peut menacer les fondements mêmes des terminologies.

> «[…] jusqu'à une date récente, la linguistique n'a fait valoir que l'évolution libre, non dirigée, de la langue. C'est l'usage effectif de cette dernière qui, dans la langue commune, sert de norme. On peut appeler cette norme la norme descriptive. En revanche, en terminologie, fertile en notions et en termes, cette évolution libre de la langue mène à une confusion inacceptable… » (Wüster, 1981, 65).

Pour Wüster en effet, la terminologie est normative, par essence et/ou, par objectif. Comme l'a montré Slodzian entre autres (Slodzian, 1995), Wüster croyait en l'existence d'une langue scientifique épurée (en tout cas épurable) de ce qu'il considérait comme les éléments qui nuisaient à une communication transparente.
La réalité de la pratique terminologique se révèle tout autre. En effet, les textes, entendus comme des productions langagières effectives, sont nécessairement pris en compte parce que les terminologues ne peuvent s'appuyer sur leurs seules intuitions linguistiques dans des domaines où ils n'ont pas de compétence. Pour contourner cette « non-compétence », les terminologues font appel à des experts qu'ils interrogent mais aussi à des productions de toutes natures : manuels, documents d'entreprises, listes de termes existantes… Les praticiens et plus encore les chercheurs se heurtent alors à une double difficulté. D'une part, cette prise en compte des réalisations possibles n'est pas systématisée ; or, il est bien évident qu'il y a une difficulté à considérer comme autant d'attestations des textes qui s'adressent à des non-experts et d'autres qui sont très spécialisés ou encore, des textes qui ont une visée principalement injonctive et d'autres qui ont une visée descriptive. D'autre part, la tâche qui est demandée au terminologue revient à construire une norme à partir d'usages attestés. Très rapidement, il est confronté au manque de balisage que constitue cette situation si elle n'est pas accompagnée d'une réflexion sur l'objectif de la normalisation.
Pourtant, nourrie par les travaux sur l'analyse de corpus en provenance de la linguistique et sur les réflexions concernant la prise en compte des applications en TAL et ingénierie des connaissances, la terminologie textuelle est devenue une problématique à part entière qui pourrait constituer un lieu de réflexion majeur sur la confrontation de la linguistique avec le réel via l'utilisation de corpus.

3.2.Des textes au réseau terminologique

Dans une vision classique de la terminologie, élaborer une terminologie à partir de textes revient à retrouver dans ces textes les termes et les relations qui les relient, « par essence ». Le réseau terminologique préexisterait et il suffirait de voir comment il se met en mots dans les textes. Ce point de vue est proche de celui d'une sémantique lexicale qui cherche à trouver dans les textes le sens des mots, qui subsume et explique toutes les occurrences. Dans le cas de la terminologie, où l'emprise avec la praxis est très importante (la plupart du temps en effet, les terminologies sont associées à des métiers ou des pratiques sociales), cette position est très problématique car elle ne tient pas compte de la créativité inhérente à la situation de dialogue. De plus en plus, la terminologie textuelle se rapproche de l'analyse de discours qui considère que toute analyse de textes est un construit et est soumis à une interprétation. Il s'agit, en l'occurrence, de construire, à partir d'usages attestés, un réseau lexical ; mais ce réseau lexical est difficilement envisageable comme LA terminologie ; en effet, ce réseau est d'abord associé à un groupe de locuteurs mais aussi à un objectif précis. On considère ainsi qu'il n'est pas équivalent de construire une terminologie pour la traduction, la recherche d'information ou la constitution d'un

thésaurus. Cet objectif interprétatif apparaît ainsi comme un élément majeur dans l'analyse d'un corpus.

Plutôt que de considérer les textes comme des attestations d'une connaissance préexistante, une manière alternative consiste à considérer le corpus comme le point de départ, voire comme la référence à l'étude. La constitution du corpus devient alors un moment crucial de l'analyse et elle doit nécessairement être précédée par une réflexion sur l'objectif de l'analyse terminologique (qui peut être théorique ou appliqué) puisque c'est cet objectif qui va guider la constitution du corpus lui-même puis l'élaboration du réseau terminologique.

Une fois ce cadre posé, on peut considérer que la construction d'une terminologie relève d'une problématique clairement linguistique que l'on peut énoncer de la manière suivante. Il s'agit de passer d'une représentation discursive, syntaxique et linéaire à une représentation spatiale et iconique dans laquelle subsiste des éléments de la représentation discursive (en l'occurrence les termes, qui peuvent avoir été « normalisés » (Rastier, 1995) *i.e*, légèrement modifiés : lemmatisés par exemple). Il s'agit donc d'introduire une rupture dans le continuum discursif pour le discrétiser et donner à certains éléments un contour défini, une définition. En effet, comme le signale A.Rey, il y a, dans le mot *terme* comme dans le mot *définition* une idée d'élément circonscrit, fini (Rey, 1979). Cette délimitation n'est pas un donné mais un construit. La question qui se pose finalement est celle de savoir sur quelles bases, en mettant en œuvre quelles connaissances va s'élaborer le réseau. Ce qui est sûr, c'est que ces connaissances vont nécessairement s'ancrer dans des éléments textuels ; il faut donc comprendre pourquoi on attribue un rôle particulier (termes ou relations) à certains éléments dans la représentation en réseau.

Le plus souvent, ce sont d'abord les termes qui sont recherchés. Dans la très grande majorité des cas, ce que l'on décide de considérer comme termes sont des noms ou des syntagmes nominaux, sans doute parce qu'on associe à la forme nominale un haut niveau de stabilité et de pouvoir de désignation. Il faut avoir fait l'expérience qui consiste à choisir les termes parmi la liste de tous les syntagmes nominaux d'un corpus (fournis par un extracteur de termes) pour se rendre compte de la difficulté qu'il peut y avoir à décider ce qu'est un terme. L'idée qu'un terme « désignerait une notion de manière univoque à l'intérieur d'un domaine » (OLF, 1985) n'est, en réalité, pas opérationnelle.

Dans ce type d'approche, qui se focalise d'abord sur le repérage des termes, les relations sont ajoutées dans un second temps.

Une approche alternative consiste à travailler d'abord sur les relations pour considérer que les termes seront les éléments qui sont mis en relation dans le corpus. Le repérage de cette mise en relation met en scène la notion de « marqueur de relation », notion beaucoup plus complexe qu'on pourrait le croire et qui interroge fondamentalement les possibilités de systématiser les analyses de discours, quelles qu'elles soient.

3.3. Les marqueurs de relations conceptuelles

La notion de marqueur de relations conceptuelles est utilisée par différentes disciplines, même si les points de vue sont légèrement différents. Ainsi Cruse parle de « diagnostic frames » (Cruse, 1986), Ahmad de « knowledge probes » (Ahmad et *al.*, 1992), Meyer de « knowledge rich contexts » (Meyer, 2000). Il s'agit d'éléments linguistiques, le plus souvent lexicaux ou lexico-syntaxiques, qui permettraient de repérer systématiquement une ou l'autre relation conceptuelle. La vision systématique souvent associée à cette notion de marqueur suppose des hypothèses fortes :

- Il y a un système linguistique unique et stable qui autorise à rechercher par introspection les marqueurs de relation.
- Les relations repérables en corpus sont essentiellement binaires. Dans une perspective terminologique, la prise en compte de relation n-aires n'est pas envisagée ; par exemple, on ne peut pas représenter l'énoncé : *l'entreprise X avait deux filiales en 1995* par deux relations co-dépendantes. Cette impossibilité est peut être due au fait que les relations n-aires se rapprochent trop d'un fonctionnement discursif (structure argumentale par exemple), ce qui est incompatible avec la vision wüstérienne.
- Les marqueurs donnent le sens de la relation. D'une certaine façon, l'apparition d'un marqueur déclencherait l'interprétation sous la forme d'une relation ; les marqueurs fonctionnant alors sur un mode indiciel plus que réellement sémantique.

La réalité des phénomènes que l'on peut observer dans les corpus oblige à revoir ou, à tout le moins, à affiner ces hypothèses (Condamines, 2002).
Ainsi, les marqueurs sont des éléments linguistiques de toutes natures (lexicaux, syntaxiques voire typographiques) auxquels on attribue un statut particulier, métalinguistique, qui donne la possibilité de leur associer une relation conceptuelle. La question est donc de savoir avec quel degré de systématicité peut s'instaurer l'attribution d'un rôle métalinguistique à certains éléments. La nature du corpus semble jouer ici un rôle majeur. Trois modes de liens entre marqueur et corpus peuvent être identifiés : pas de dépendance, dépendance totale, dépendance avec le *genre* du corpus.

Pas de dépendance entre le marqueur et le corpus
Ces cas, qui sont ceux qui sont décrits le plus souvent dans les travaux sur les marqueurs, sont en réalité assez peu fréquents. Il s'agit de marqueurs que les auteurs retrouvent par introspection et qui sont présentés comme généraux ; ils sont aussi associés à des relations considérées comme générales : relation générique ou méronymique, par exemple : [déterminant N1 comme déterminant N2] (*Un département comme la Seine bénéficie à la fois d'arrivées d'enfants et de scolaires)*
La systématicité de ces marqueurs est toutefois à relativiser. D'une part parce qu'ils ne sont pas toujours utilisés dans les textes, d'autre part parce qu'ils peuvent renvoyer à une autre relation ; ainsi ce même marqueur peut renvoyer à une relation de « comparaison », par exemple

: *On comprend que les lycées professionnels et d'enseignements général, comme l'université, soient très peu tournés vers les formations scientifiques et technologiques (*cet extrait provenant du même corpus que l'extrait précédent).

Dépendance totale entre le marqueur et le corpus
Certains marqueurs sont imprédictibles par introspection et ne se révèlent qu'au cours de l'analyse, souvent approfondie, d'un corpus. Ainsi, dans un manuel de génie logiciel fourni par EDF, nous avons identifié le marqueur suivant pour la relation de condition :

[((phase, étape), déverbal) + (lorsque, dès que) + V au passif]

comme dans :
La phase d'intégration du composant peut commencer lorsque l'ensemble des éléments logiciels ont été codés.
Cette phrase doit être interprétée comme : *la phase d'intégration du composant ne peut commencer que lorsque les éléments logiciels ont été codés*. Il s'agit donc bien d'une condition et son marqueur était imprédictible par introspection (même s'il est identifiable lors de l'analyse). Ce type de marqueur fonctionne plutôt sur le mode épilinguistique, au sens culiolien d'activité métalinguistique inconsciente.

Dépendance entre le marqueur et le genre du corpus
Certains éléments fonctionnent comme marqueurs principalement dans des corpus qui relèvent d'un genre précis.
Prenons le cas de *chez* et *avec.*
Dans certains cas, *chez* peut être utilisé dans des phrases où s'exprime une relation de méronymie (Condamines, 2000), par exemple :

Chez les primates, la mandibule a des mouvements verticaux

Mais cette possibilité est nettement plus présente dans des corpus d'un genre bien précis : les textes doivent relever des sciences naturelles et être didactiques (ce fonctionnement est ainsi moins présent dans des quotidiens généralistes que dans l'*Encyclopedia Universalis*, aux rubriques relevant des sciences naturelles).

Tout comme *chez*, *avec* peut « marquer » la méronymie mais dans des conditions bien précises. En effet, l'analyse de corpus montre que, en tout cas pour la valeur méronymique de *avec* (relation composant-tout), il est particulièrement éclairant d'introduire, dans la description, la notion de genre textuel. En effet, cette valeur n'apparaît pas de manière similaire dans tous les corpus : elle est très rare dans les corpus littéraires (seulement 6,4 % des occurrences de *avec* dans Germinal, par exemple : *il était petit, le cou énorme, les mollets et les talons en dehors, avec de longs bras*)
et, au contraire, très fréquente dans certains textes, par exemple :

– les catalogues de jouets (68,2 % des occurrences dans un catalogue de jouets):

  Château fort sur rocher avec personnage et cavaliers.

– les petites annonces immobilières (76,2 % dans un corpus de petites annonces):

A l'étage : 3 chambres avec placards.

Dans ce genre de corpus, la notion de méronymie doit elle-même être affinée. Ainsi, dans les petites annonces par exemple, elle est combinée avec la mise en valeur du logement à vendre : en témoigne l'utilisation du signe + dans les mêmes contextes qu'*avec* :

Villa […] avec hall de nuit + placards et penderie.

Par ailleurs, dans ce genre de discours, un type de méronymie peut être construit, qui doit être distinguée d'une méronymie que l'on pourrait considérer comme « ontologique » ; ainsi dans l'exemple suivant, qui met sur le même plan des éléments *a priori* bien différents :

Villa T4 avec piscine, gardien, garage, jardin, parquets

Enfin, dans les catalogues de jouets, la méronymie comporte souvent une valeur mimétique (le jouet doit ressembler à l'original destiné aux adultes) et le jouet est présenté comme contenant toutes les parties caractéristiques de l'original :

Bloc de cuisson avec hotte, four et plaque de cuisson.

Les quelques exemples que nous avons proposés mettent en évidence plusieurs caractéristiques :

– Le fonctionnement précis des « marqueurs » est difficilement envisageable sur la seule base de l'introspection.
– Les éléments appelés marqueurs ne donnent pas le sens de la relation ; ils apparaissent plutôt comme des éléments déclencheurs d'une éventuelle interprétation relationnelle. Les structures syntaxiques dans lesquels ils apparaissent peuvent constituer une manière de sélectionner les énoncés pertinents par rapport à une relation mais elles ne sont souvent pas suffisantes pour garantir l'interprétation.
– Le genre textuel joue un rôle fondamental dans l'attribution d'une possibilité d'interprétation relationnelle à une structure.

Ainsi, à travers l'étude de ce qui aurait pu relever d'une problématique uniquement appliquée (construire une terminologie), des descriptions très fines apparaissent nécessaires si l'on veut mettre en œuvre une approche systématique. Ces descriptions relèvent à part entière d'une sémantique de corpus, c'est-à-dire d'une sémantique qui s'intéresse au fonctionnement du sens dans un corpus et aux modes d'élaboration d'une interprétation. La question particulière de la terminologie textuelle permet ainsi de mener une réflexion poussée sur les possibilités de baliser l'interprétation d'un texte. Deux éléments semblent majeurs dans cette perspective : le genre textuel d'une part et l'objectif d'interprétation d'autre part, qu'il soit théorique ou appliqué. Ainsi, interpréter un texte pour construire une terminologie, pour analyser le fonctionnement des adverbes ou des déterminants, identifier les thèmes récurrents d'un texte, vérifier l'hypothèse de l'utilisation massive d'une structure donnée, pour ne prendre que quelques exemples, sont des objectifs qui, non seulement guident l'interprétation, mais aussi doivent guider la constitution du corpus.

4. Conclusion

La prise en compte du « réel » en linguistique est une question souvent posée (Siblot, 1990). La sémantique cognitive associe cette prise en compte à la perception, considérée comme proche d'un locuteur à l'autre, et qui expliquerait les universaux de langage. Utiliser les corpus comme base d'étude pour la linguistique revient à intégrer le réel d'une tout autre façon. D'une part, si, à la suite de Bakhtine (Bakhtine,1984), on peut considérer toute production textuelle comme s'inscrivant dans un dialogue, le linguiste ne peut faire l'impasse sur cette situation de dialogue, c'est-à-dire sur le contexte de production du texte et les possibilités de la caractériser. D'autre part, tout comme la production d'un texte est située, son interprétation l'est aussi. Cela est clair et évident pour la terminologie textuelle qui est souvent associée à une pratique et à des besoins identifiées. Mais ce qui *a priori* pourrait apparaître comme une situation d'interprétation sémantique non-standard pourrait bien être un cas particulier de ce qu'est toujours une analyse sémantique de corpus : une construction basée sur une interprétation. On peut espérer comprendre sur quelles bases se fait l'interprétation en faisant intervenir la situation de production des textes et la situation d'interprétation. Ainsi, la terminologie textuelle, peut-être parce qu'elle aborde une problématique éminemment interdisciplinaire, pourrait jouer un rôle moteur dans les études qui commencent à se développer en sémantique de corpus.

Bibliographie

AHMAD K., FULFORD H., 1992, *Knowledge Processing : Semantic Relations and their Use in Elaborating Terminology*, Computing Sciences Report CS-92-Guildford, University of Surrey.

AUROUX S., 1998, La raison, le langage et les normes, Paris, PUF.

BACHIMONT B., 2000, « Engagement sémantique et engagement ontologique : conception et réalisation d'ontologie en ingénierie des connaissances », J.Charlet, M.Zacklad, G.Kassel, D.Bourigault (eds), *Ingénierie des Connaissances, Evolution récentes et nouveaux défis*, Paris, Eyrolles, pp. 305-324.

BAKHTINE M., 1984, *Esthétique de la création verbale*, Paris, Gallimard, Tel.

BIBER D., 1988, *Variation Across Speech and Writing,* Cambridge University Press.

BOURIGAULT, D. ET JACQUEMIN, C., 2000, « Construction de ressources terminologiques », J-M. Pierrel (ed), *Ingénierie des langues*, Hermès, Paris.

CONDAMINES A., 2000, « *Chez* dans un corpus de sciences naturelles : un marqueur de méronymie ? », *Cahiers de Lexicologie* n° 77, pp. 165-187.

CONDAMINES A., 2002, « Corpus Analysis and Conceptual Relation Patterns », *Terminology*, volume 8 number 1, pp. 141-162.

CONDAMINES A., AMSILI P., 1993, « Terminology between Language and Knowledge: an example of Terminological Knowledge Base », K.-D. Schmitz (ed), *TKE 93* Terminology and Knowledge Engineering, Frankfurt : Indeks Verlag, pp. 316-323.

CORBIN P., 1980, « De la production des données en linguistique introspective », A.M. Desseaux Berthoneau (ed), *Théories linguistiques et traditions grammaticales,* Lille, PUL, pp. 121-179.

CRUSE D.A., 1986, *Lexical Semantics,* Cambridge, Cambridge University Press.

DACHELET R., 1994, *Sur la notion de sous-langage*, Thèse en sciences du Langage, Université paris VIII.

DUCROT O., 1980, *Les mots du discours,* Paris, Editions de Minuit.

FOUCAULT M., 1966, *Les mots et les choses*, Paris, Tel, Gallimard.

HABERT B., NAZARENKO A., SALEM A., 1997, *Les linguistiques de corpus*, Paris, Armand Colin.

HAMON T., NAZARENKO A., (EDS), 2002, *Structuration de terminologie*, TAL (Traitement Automatique de la Langue), volume 43 – n°1/2002.

KENNEDY G., 1998, *An introduction to Corpus Linguistics*, London and New York, Longman.

KLEIBER G., 1999, *Problèmes de sémantique, la polysémie en question*, Paris, Villeneuve d'Asq, Presses Universitaires du Septentrion.

MEYER I., 2000, « Extracting Knowledge-rich Contexts for Terminography : A Conceptual and methodological Framework », D.Bourigault, M.C. L'homme, C.Jacquemin (eds), *Recent Advances in Computational Terminology*, John Benjamins. pp. 279-302.

MEYER I., BOWKER L., ECK K., 1992, «Cogniterm: An Experiment in Building a Terminological Knowledge Base». *Proceedings 5th EURALEX International Congress on Lexicography*, Tampere, Finland.

OLF, 1985 : Vocabulaire systématique de la terminologie. Montréal, OLF, 1985.

PEARSON J., 1998, *Terms in Context,* Amsterdam and Philadelphia, John Benjamins.

RASTIER F., 1995, « Le terme : Entre ontologie et Linguistique », *La Banque des Mots* n°7, Numéro spécial, pp. 35-64.

RASTIER F., 2001, *Arts et Sciences du texte*, Paris, PUF, formes sémiotiques.

REY A., 1979, *La terminologie, noms et notions*, Paris, PUF, Que sais-je ?.

SIBLOT P., 1990, « Une linguistique qui n'a plus peur du réel ». Cahiers de Praxématique n°15. pp.12-36.

SINCLAIR J., 1995, Collins Cobulid Dictionary of Idioms.

SLODZIAN M., 1995, « La doctrine terminologique, nouvelle théorie du signe au carrefour de l'universalisme et du logicisme », *ALFA (Actes de Langue Française et de Linguistique), Terminologie et langues de spécialité*, 7/8, Dalhousiana, Halifax, Nova Scotia, Canada, 1994-1995, pp. 121-136.

WUSTER E., 1981, « L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses », G.Rondeau et H.Felber (eds), *Textes choisis de terminologie*, GIRSTERM, Université de Laval, Québec, pp. 55-108.